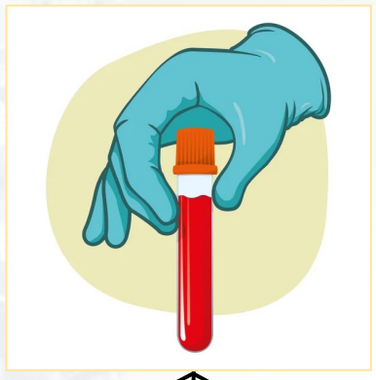
Flexible pipeline for identification reliable genetic markers in large genomic datasets

Gap.Jumper

Pawel Rosikiewicz
Data Scientist, Pipeline Developer

False Positive Leakage in BIG DATA



EVEN IF WE USE ONE BLOOD SAMPLE

WE WILL STILL GET DIFFERENT RESULS FROM TIME TO TIME (..... ATGC....)

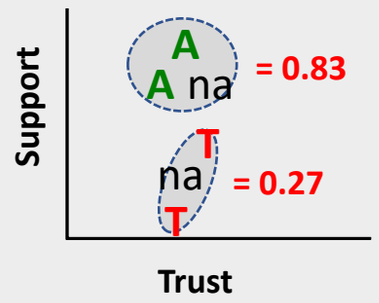
It becomes a major problem when:

- 👉 DETECTING RARE MUTATIONS
- 👉 USE LOWER QUALITY DATA
- 👉 CAN NOT APPLY IMPUTAITON

My Solution

I IMPLEMENTED APPROACH USED FOR NAVIGATION IN DRONES & AIRPLANES

GapJumper calculates Evidence Scores based only on empirical data



New algorithm

- ✓ Kalman Filter
- ✓ Semi-Supervized
- ✓ DST-Based
- ✓ No Imputation

NEW POSIBILITIES FOR

- 👉 SCREENING STUDDIES WITH LOW-COST SEQUENICNG METHODS
- 👉 QC & SEQUENCE DATA INTEGRATION

open source

GapJumper on GitHub

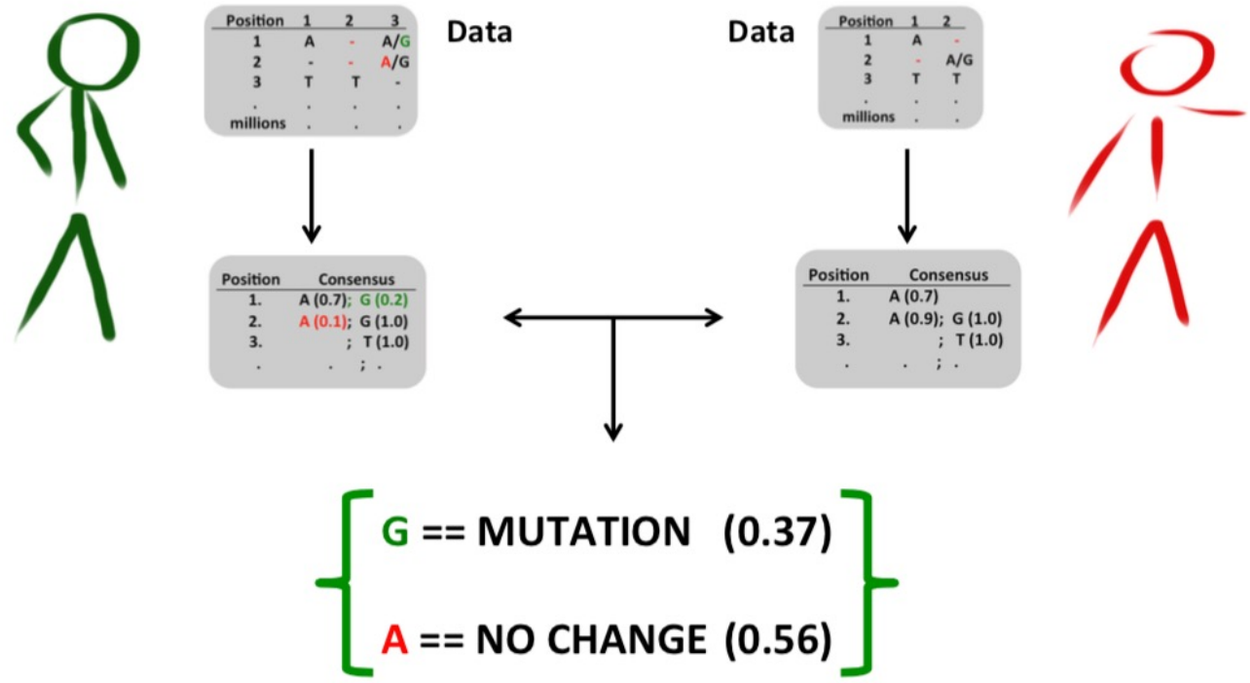
https://github.com/PawelRosikiewicz/DeepGenome

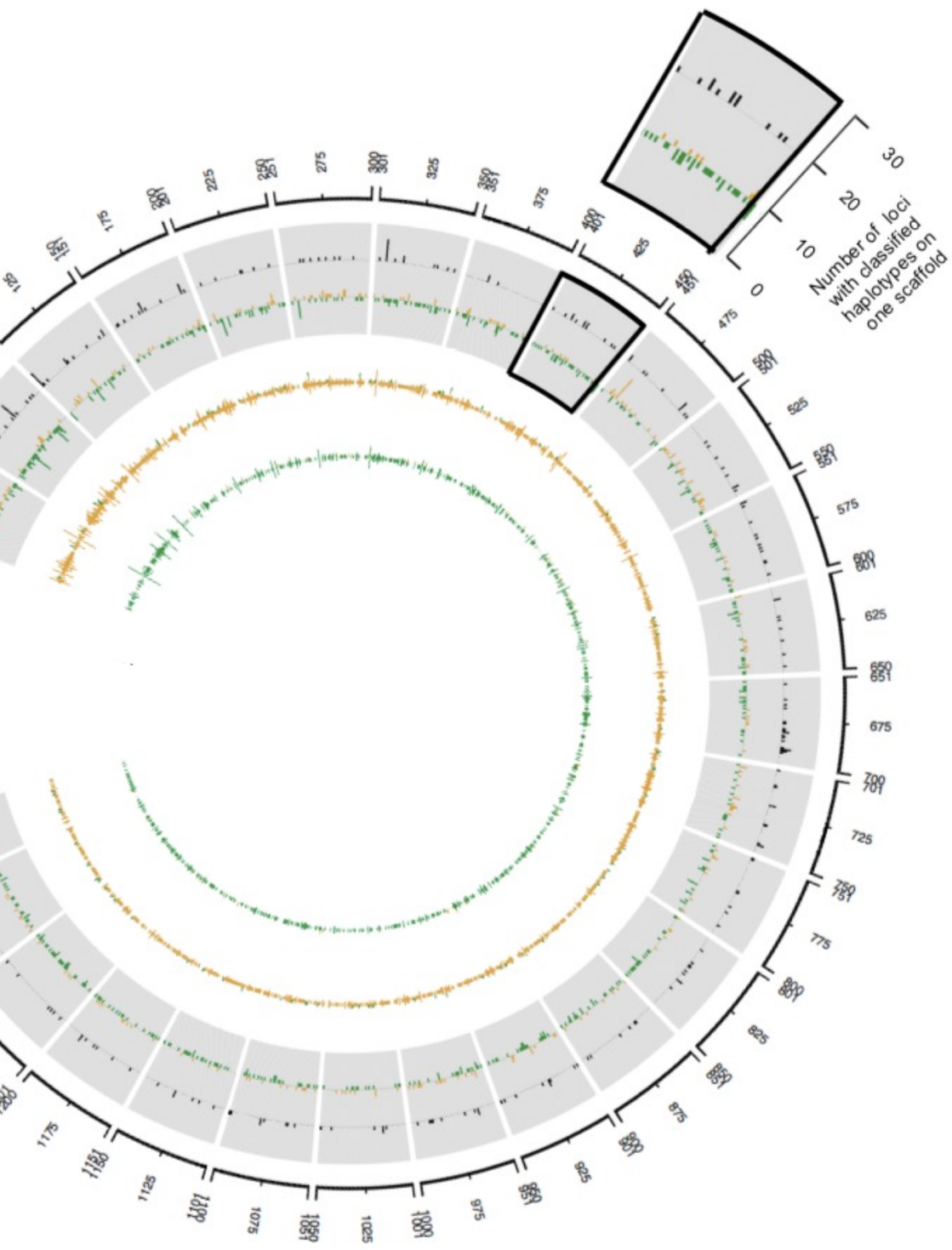
Finds reliable mutations in sequence data

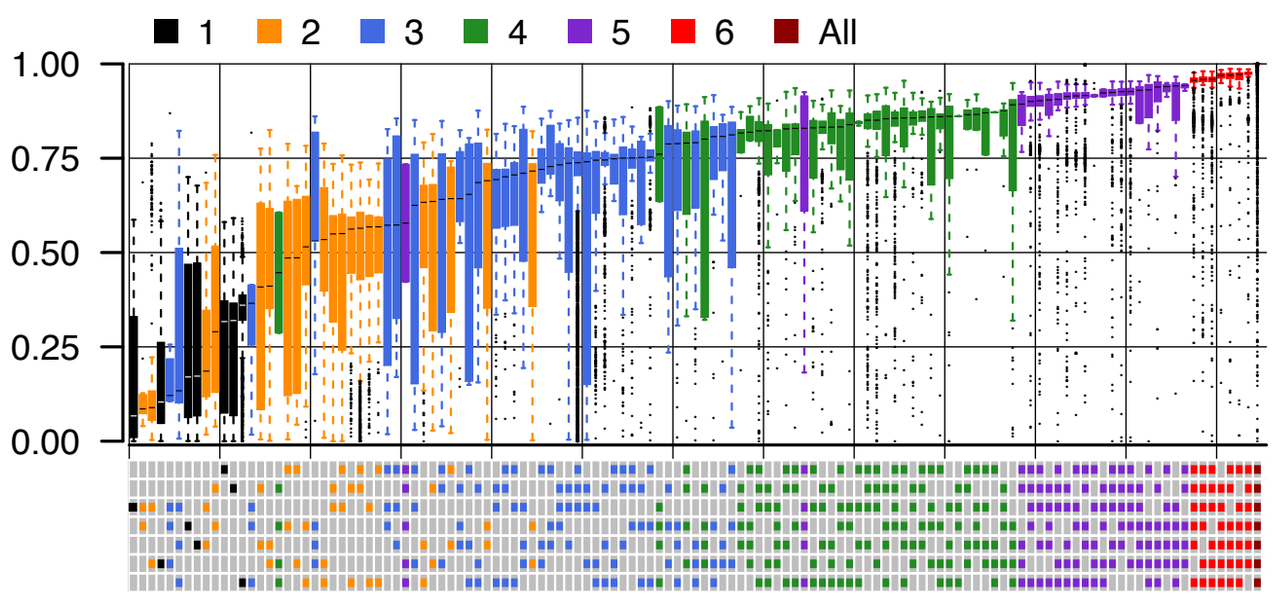
Gap.Jumper v.01

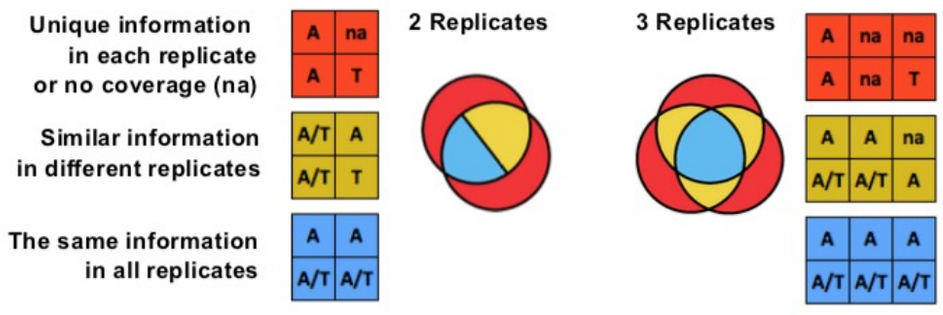
Patient

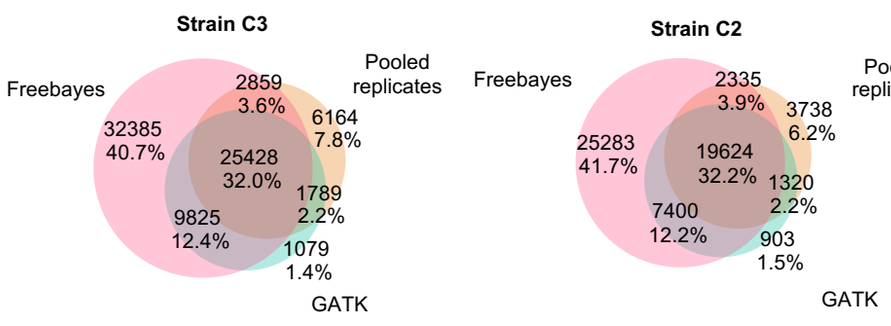
Control

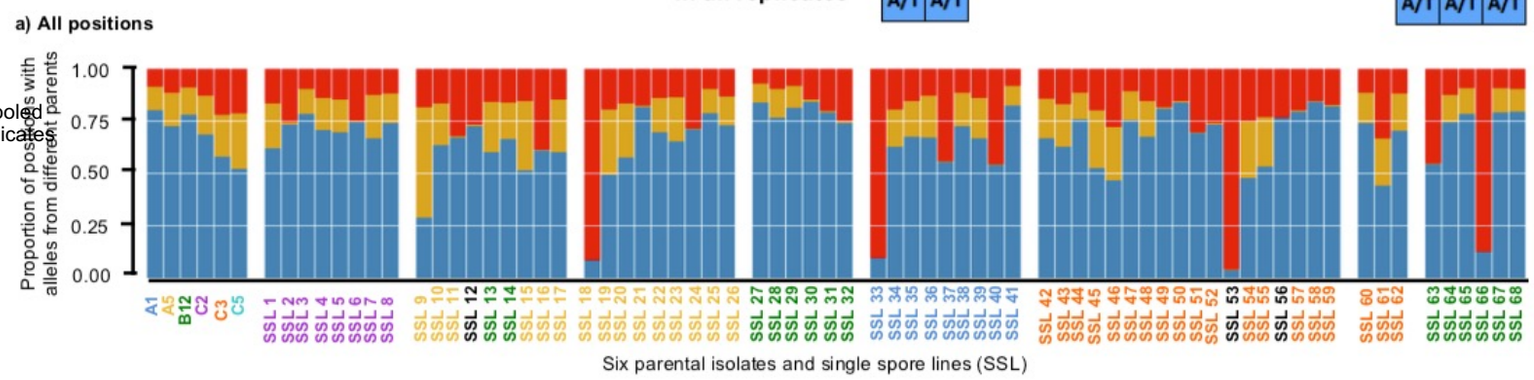




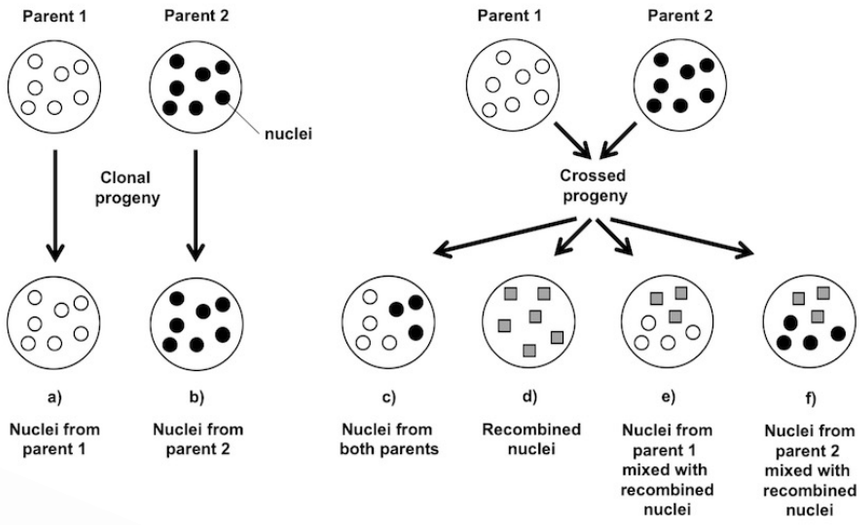


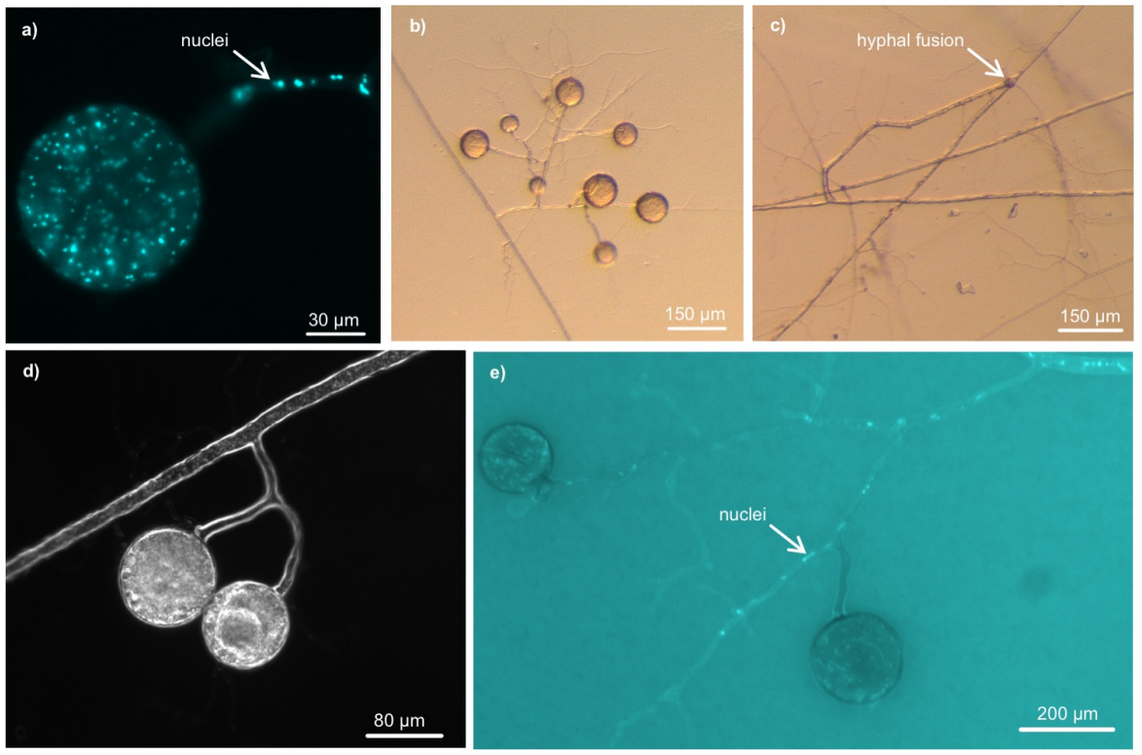


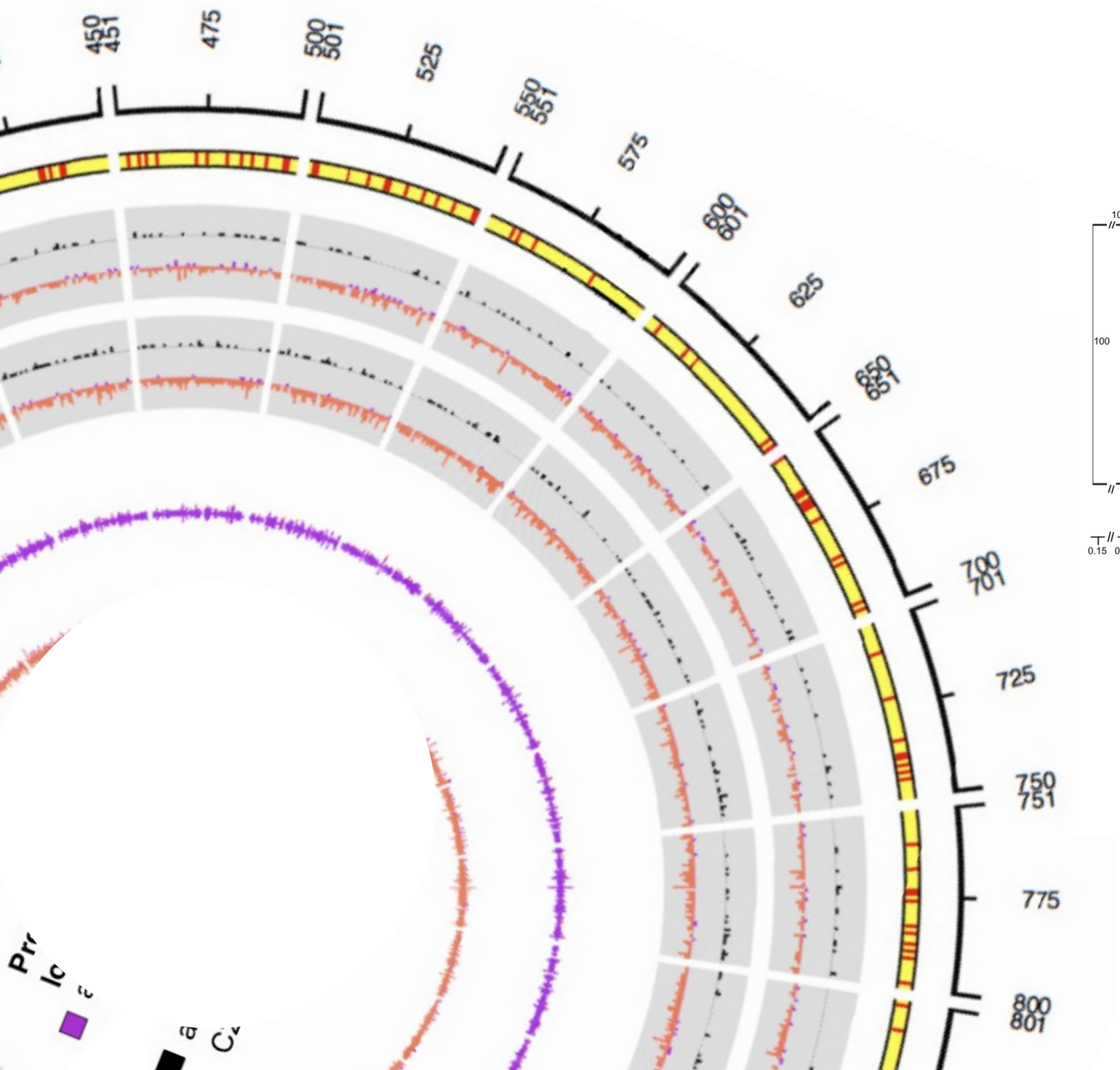


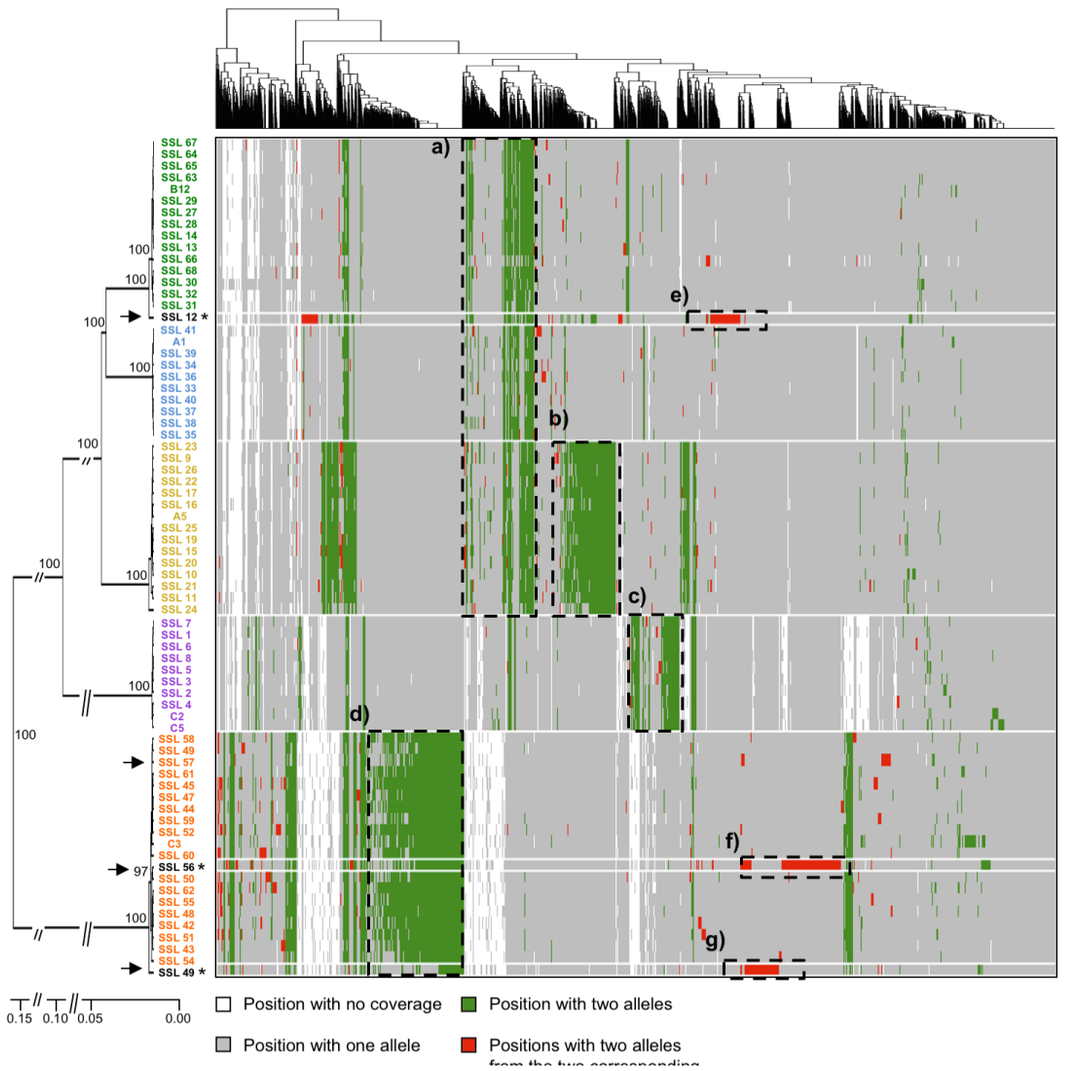


Production of new Strains Biofertilizer Fungi









Six different models used to construct simulated data (I- VI)

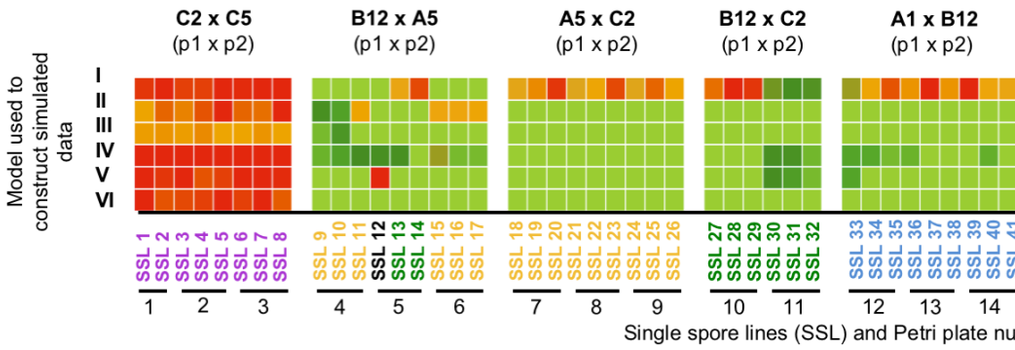
Clonal lines

- **I** Clone of parent 1 (p1)
- **II** Clone of parent 2 (p2)

Crossed lines

- **III** Admixture of nuclei from both parents
- **IV** Recombined nuclei
- **V** Recombined nuclei mixed with nuclei fro
- **VI** Recombined nuclei mixed with nuclei fro

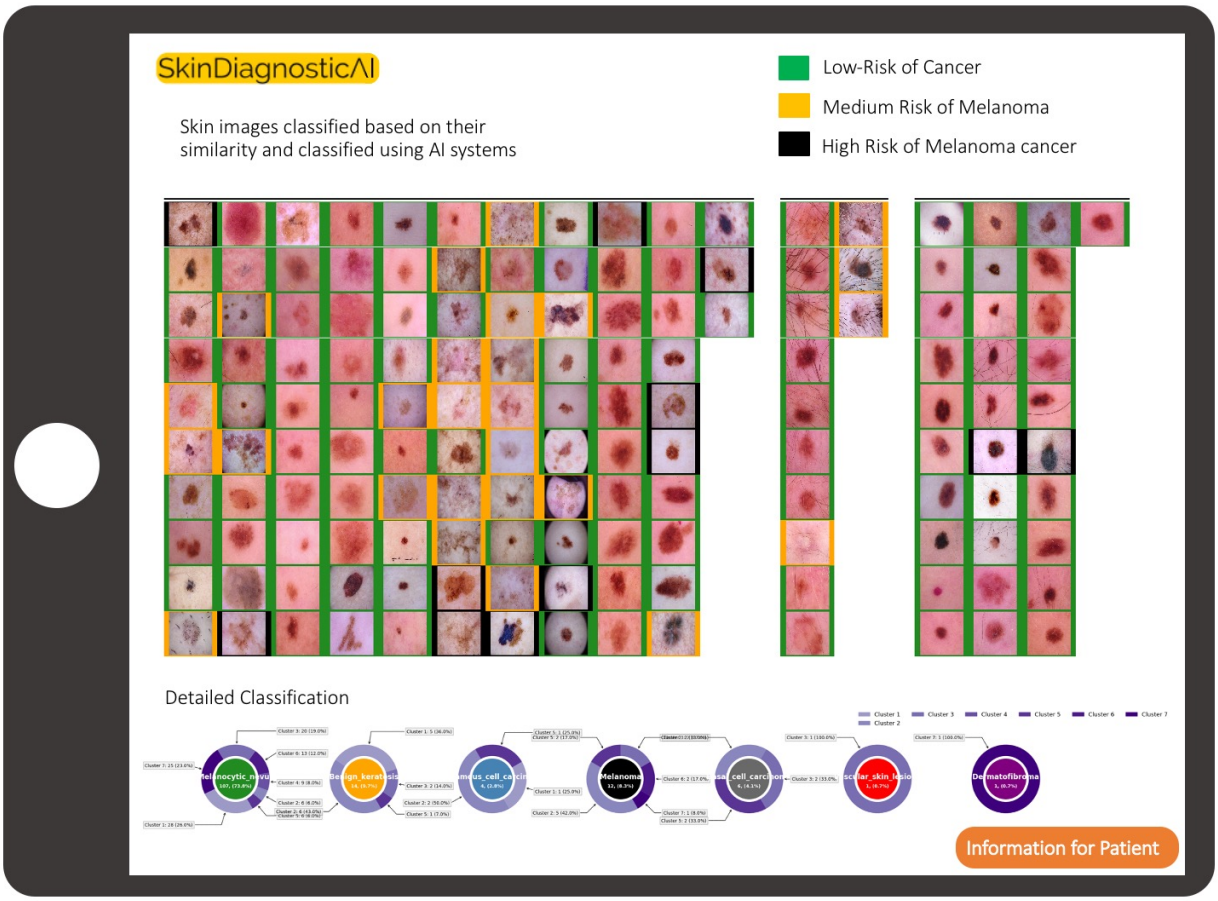
a) Simulated data compared to empirical data at all positions



SkinDiagnosticAI

Skin Cancer Detection with low-cost devices & AI

MY ROLE: Product Owner, Data Scientist



Technical goals

Define MVP requirements
What product to build?

Deploy it on the could
What model & data to use?

Improve the design
What features are important?

Build The Product
How to do it fast?

MVP SkinDiagnosticAI

PYclass Jul.2021
Add New features
And different use case

Jun.2021
Test new data sources SkinDiagnosticAI

PYclass May.2021
Partnering with other
Medical AI Groups & Teams

May.2021
Silicon Valley AI Meetup SkinDiagnosticAI

Current Moment

Apr.2021
- AI Readiness Assessment
- Re-Define Value Proposition

Mar.2021
Adopting Renku Platform

PYclass Feb.2021
MVP – light version

Data Science Eng.
Oksana

Jan.2021
New Team members

Data Sc. Manager
Onur

Nov.2020
OFFICIAL FEASIBILITY STUDY
SkinDiagnosticAI
- find the right data

HARVARD Dataverse

Sep. 2020
Tools Development
on simple dataset

PRODUCT IDEA

Business Goals

Engage Users & New Developers
What do they expect?
What features to implement?

Evaluate Resources
How much it will cost?
How long it will take?

Engage specialists
How to generate the value?

Provide Proof of concept
Is that possible for my team?

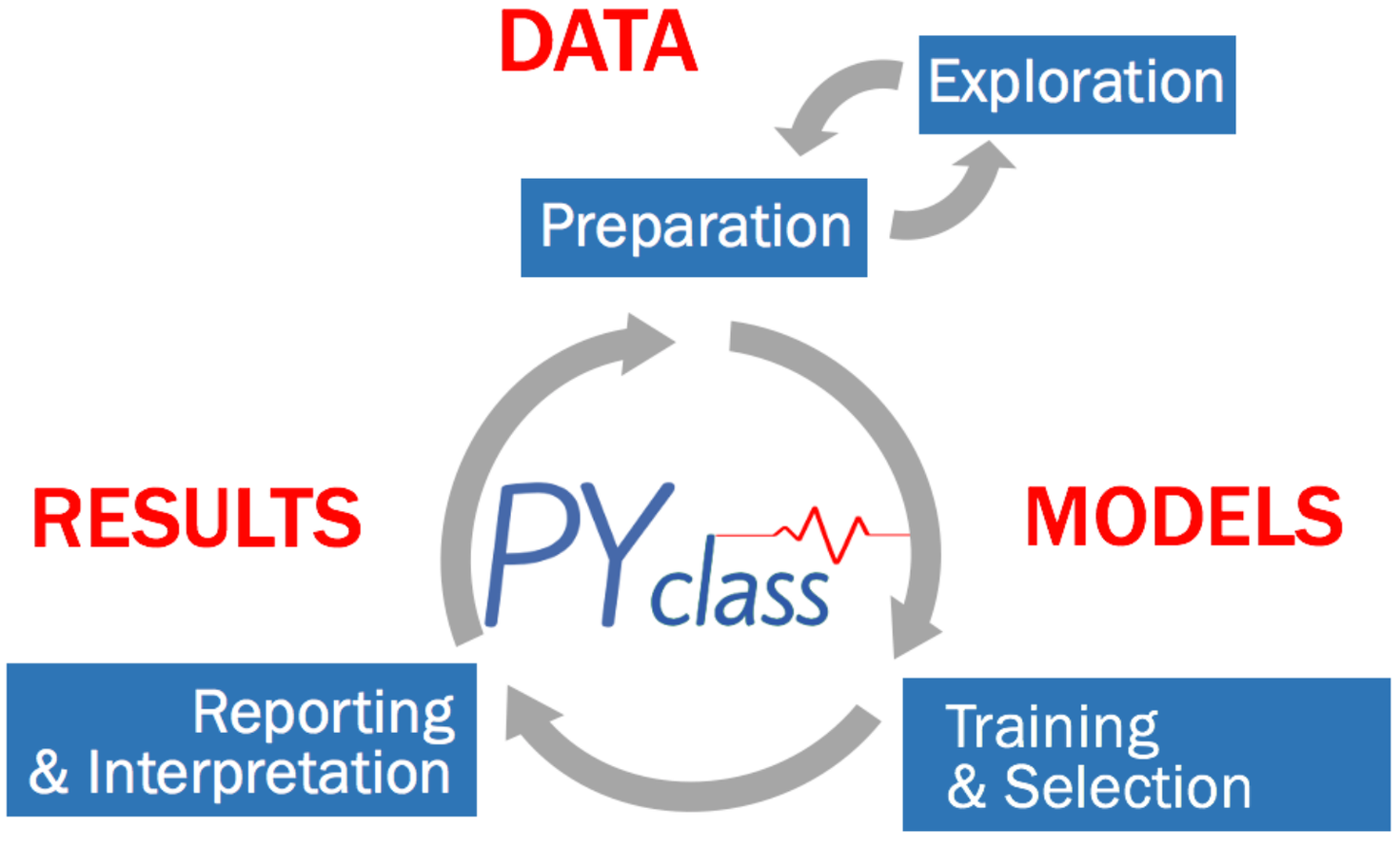
LINKS TO MY WORK

- ✓ Project Website: https://simpleai.ch/skin-diagnostic-ai/
- ✓ GitHub: https://github.com/PawelRosikiewicz/SkinDiagnosticAI
- ✓ My Talk on Meetup in Silicon Valley: https://youtu.be/W624gdkDqRQ?t=1259

PY class

AI workbench for Analysis of Medical Images

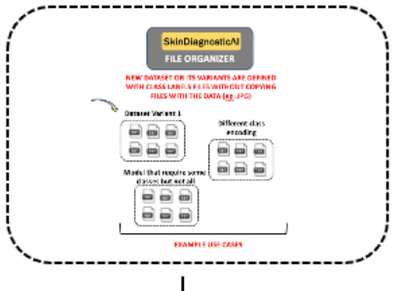
Easy interpretation



IMPUT DATA

FITERING
IMG. AUGMENTATION
BATCH SELETCTION

DATASET VARINATS



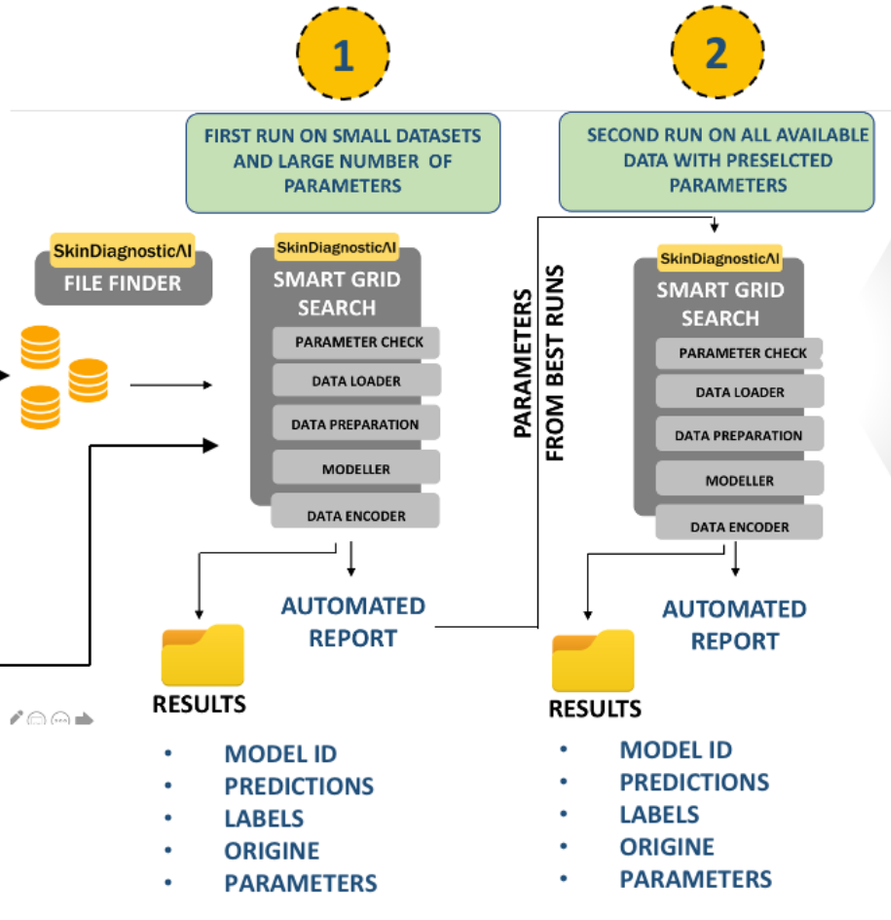
METADATA TRAVELS WITH THE FILES

Feature Extraction

Keras ImageGenerators

Medical Images classification

Visualize Trends & Examples



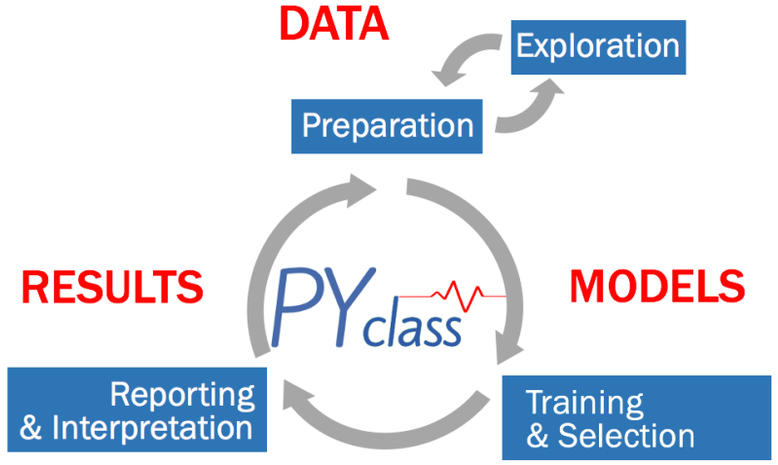
Knn
Rand Forest
SVM
Logreg

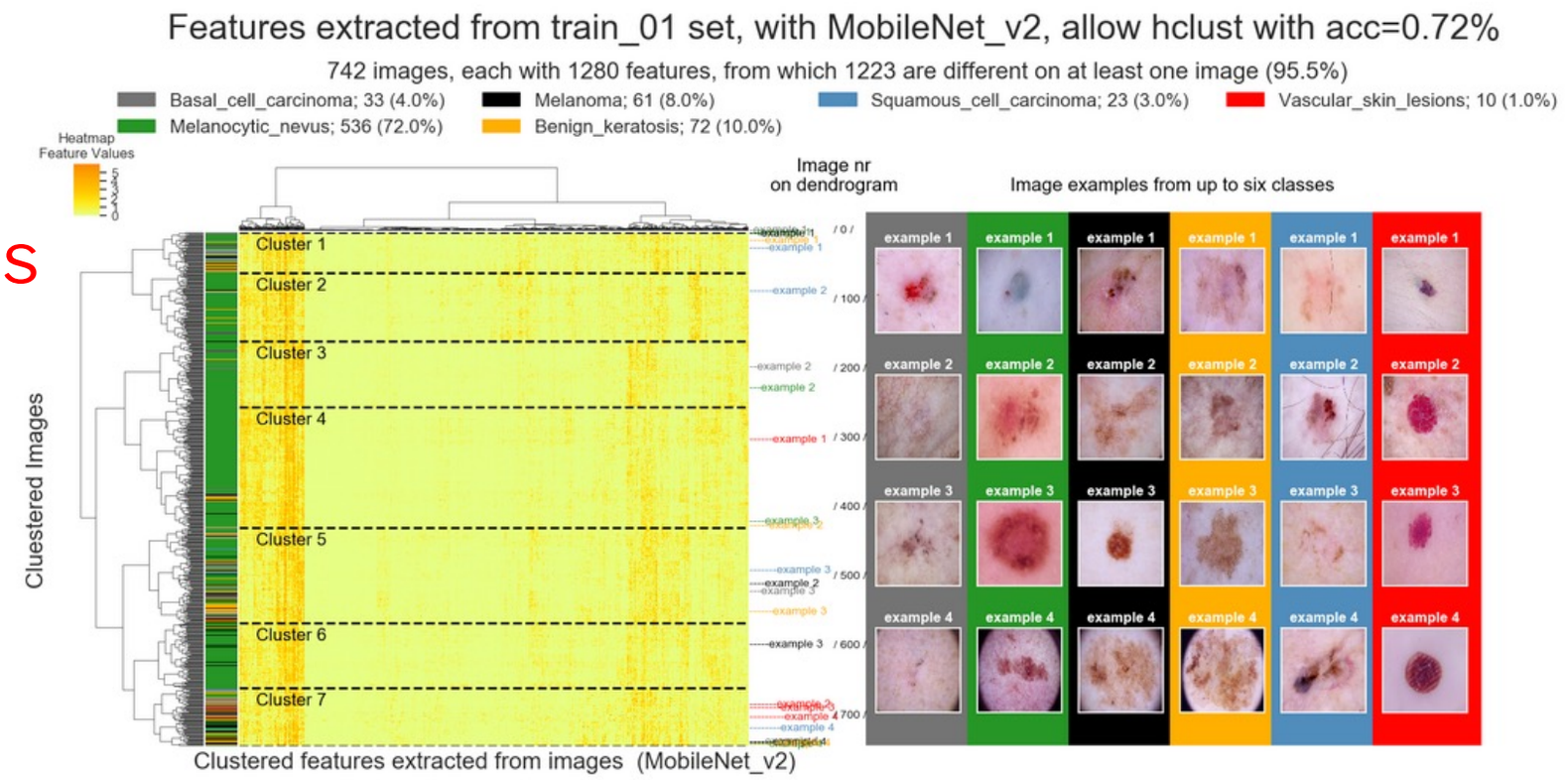
Transfer learning dense network

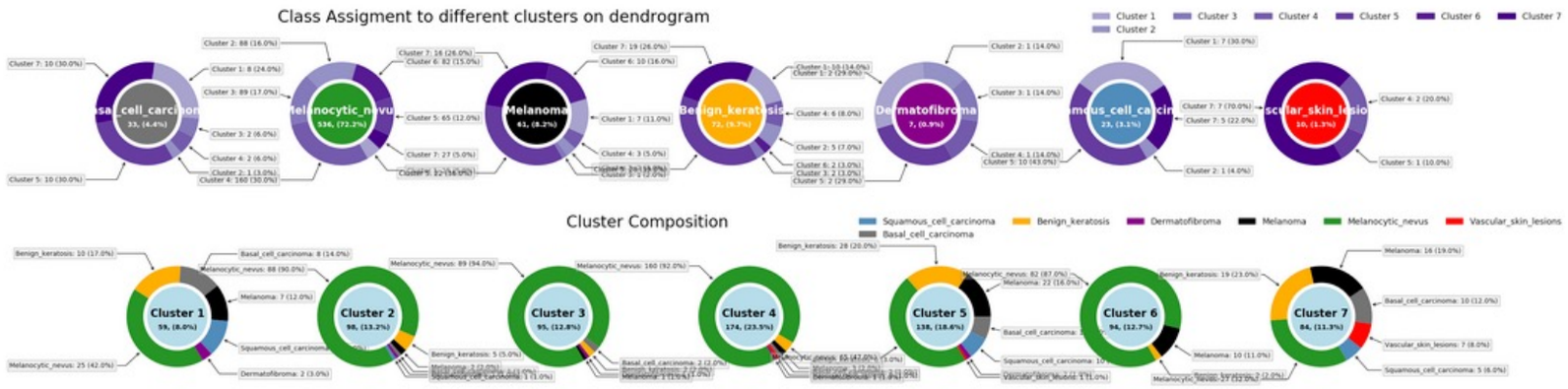
De novo ConvNets different NN arch.

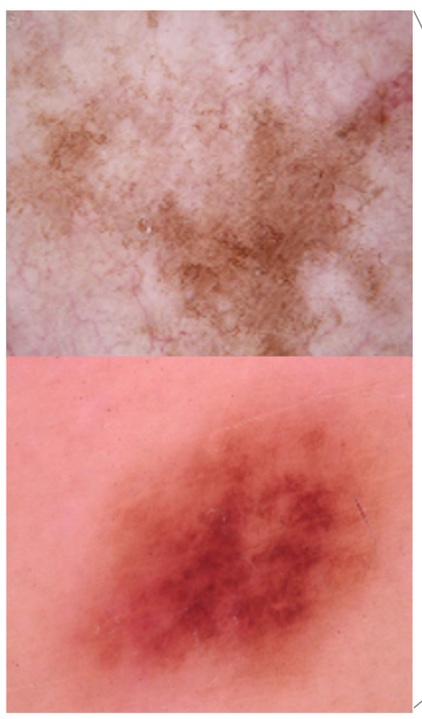
PYclass

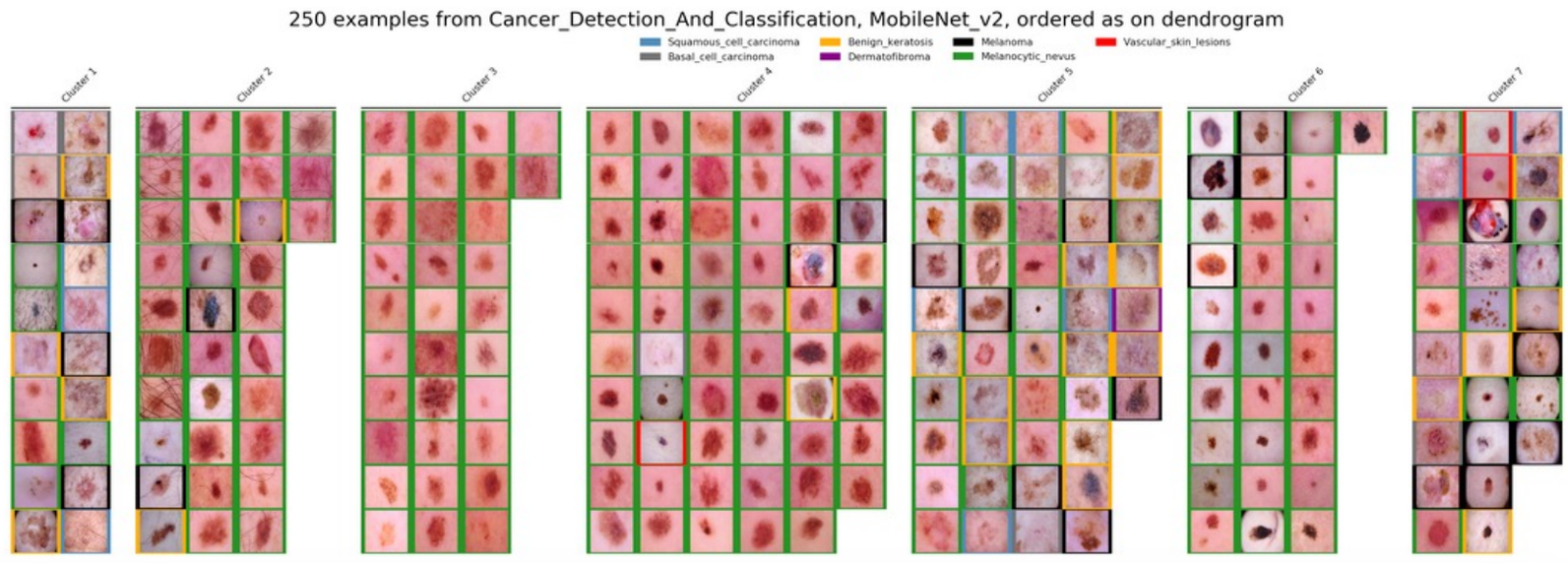
AI workbench for Analysis of Medical Images

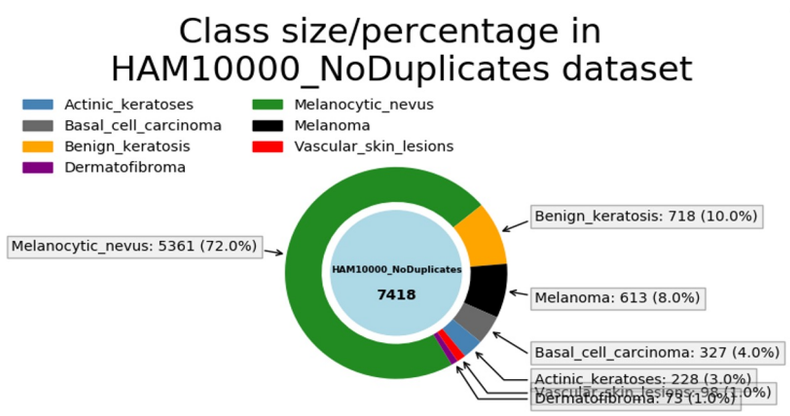


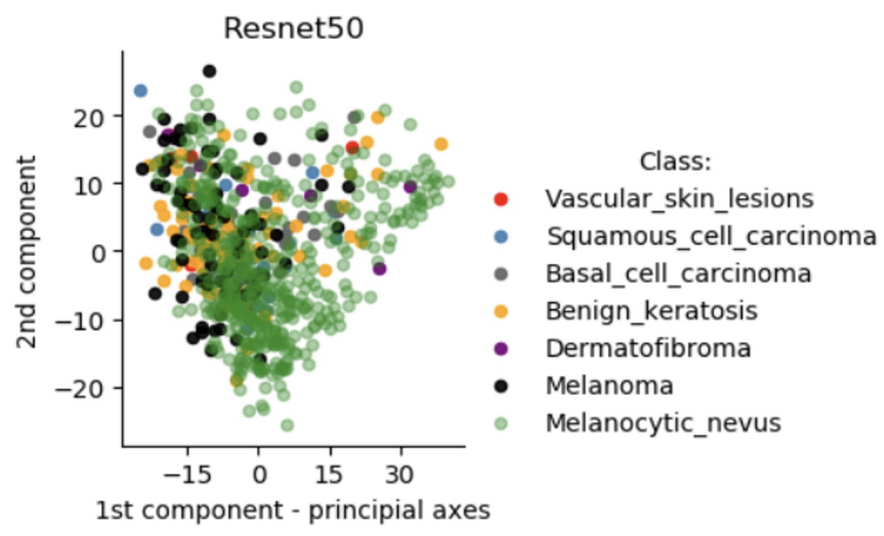






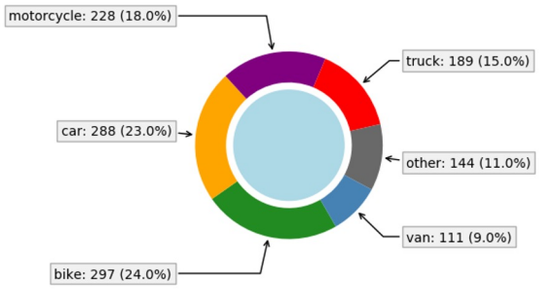


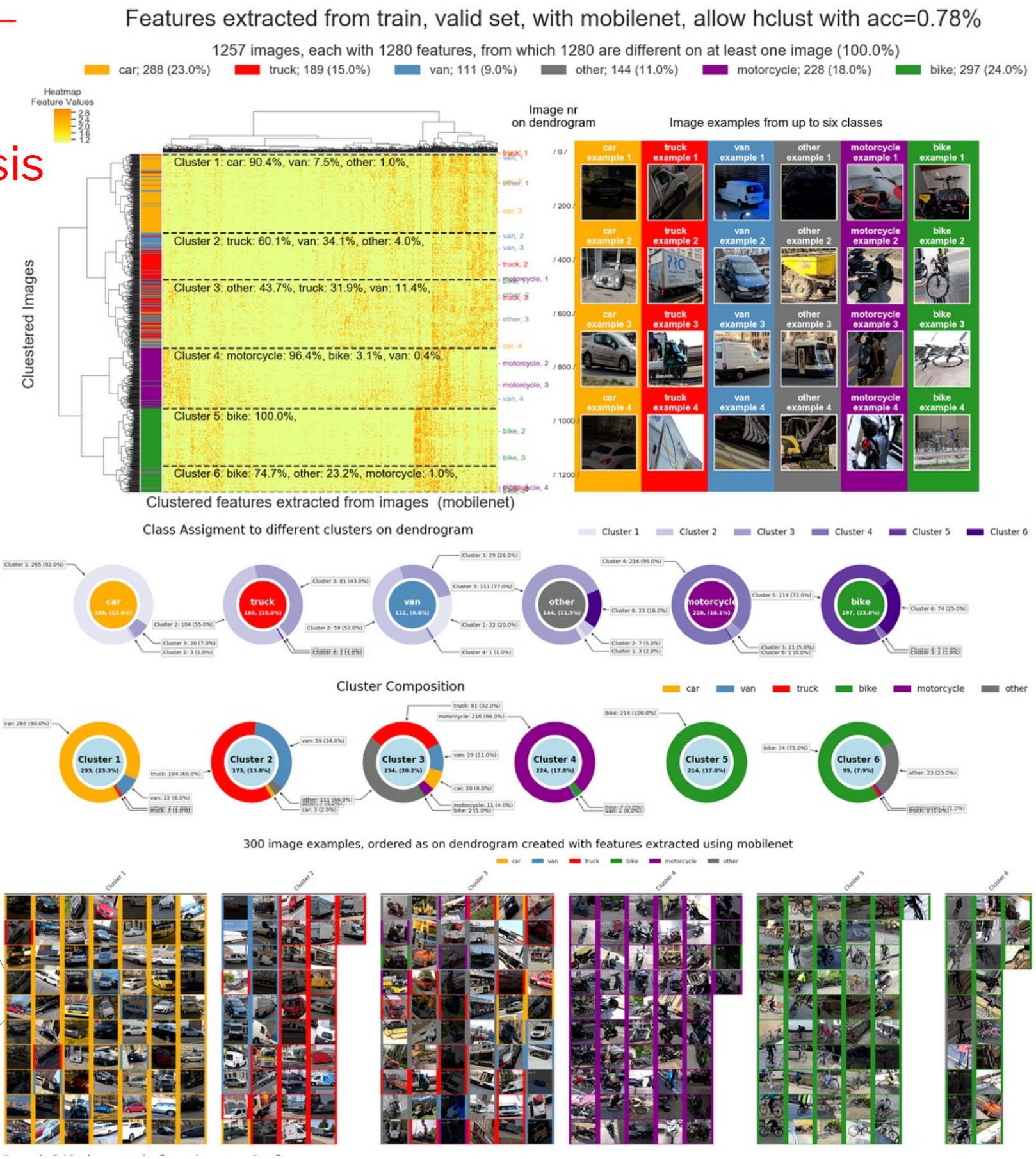




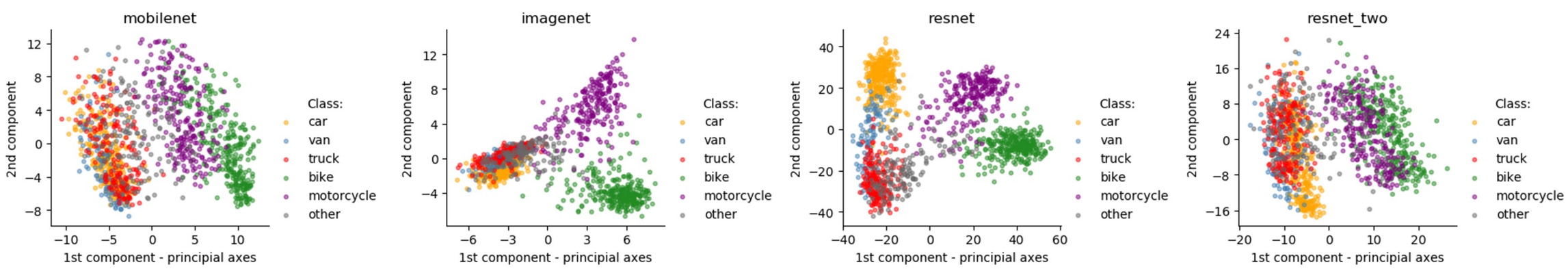
PY class

AI workbench for Analysis of Medical Images



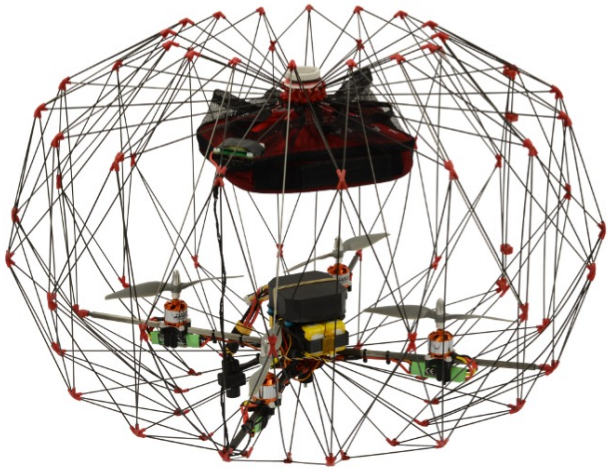


PCA: visualization of extracted features using the first two pricipial components

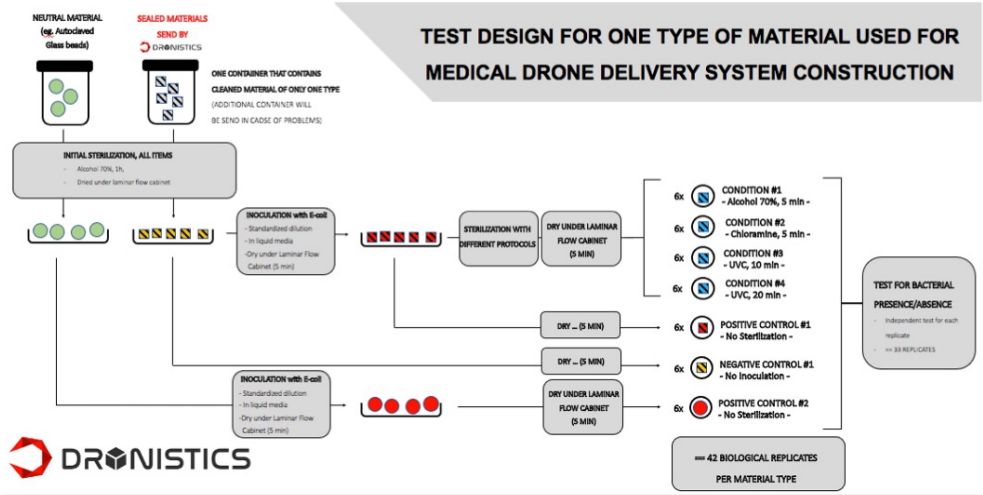


Project Manager @ DRONISTICS

Medical Delivery Drones



Experiment Design | Project & Team Management | Data Analysis



MY WORK

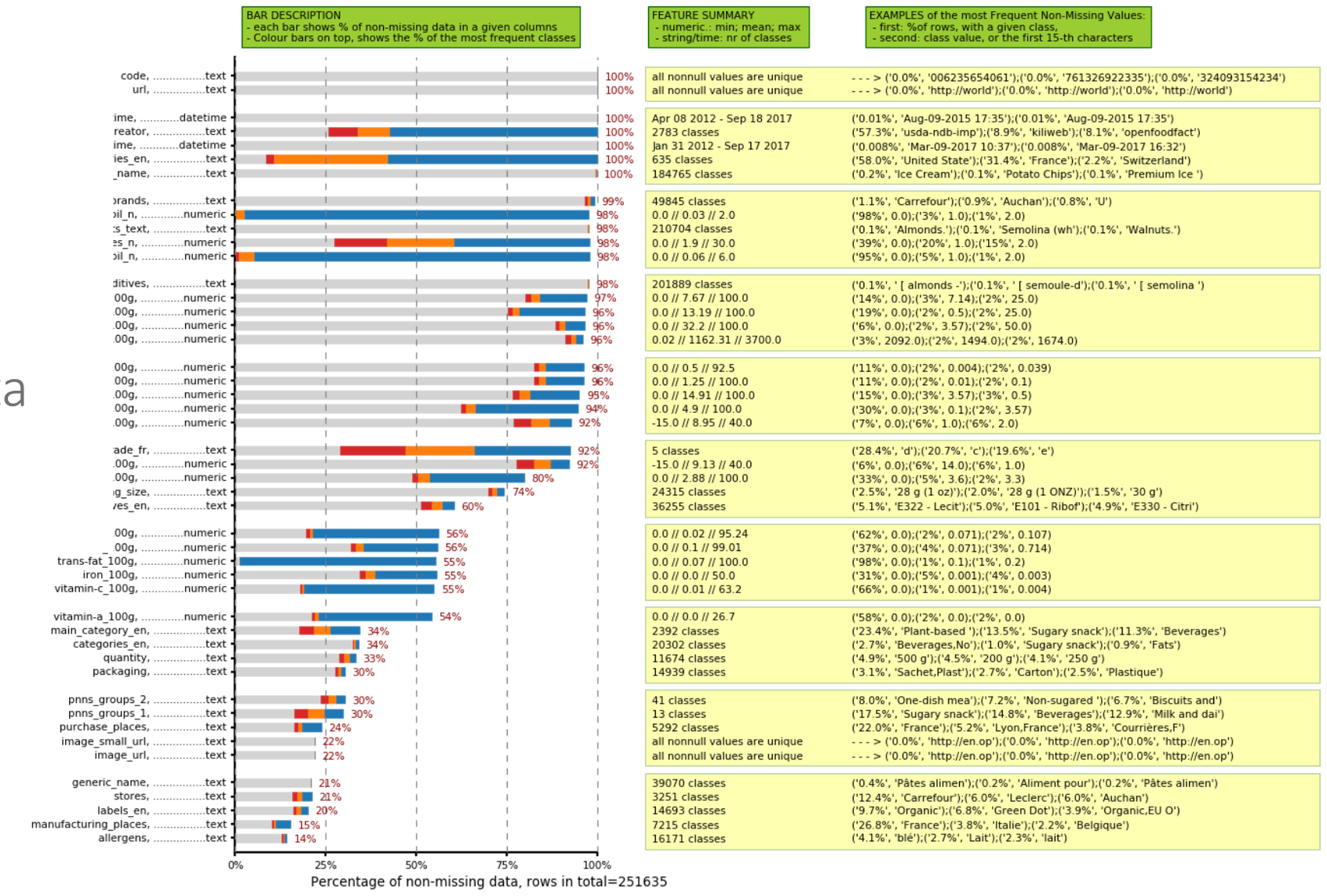
I led R&D team @ King's College London on adapting drone maintenance procedures to requirement from hospitals, & rescue services.

Data Frame Explorer

My open-source library for fast exploration, analysis, and cleaning of Python Data Frames

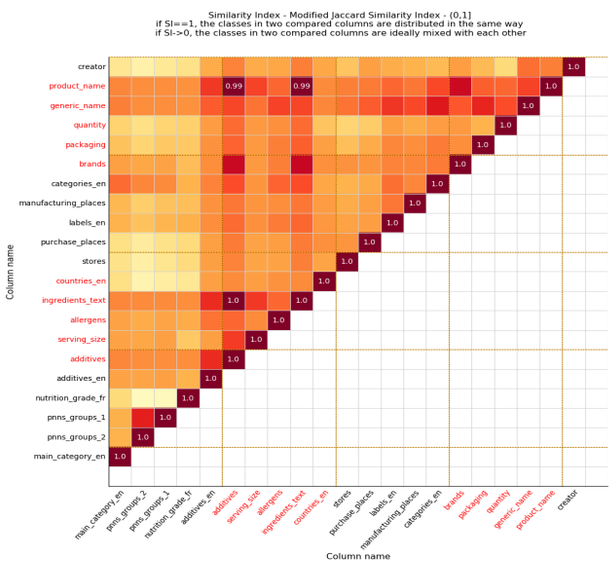
Fast summary for hundreds of features

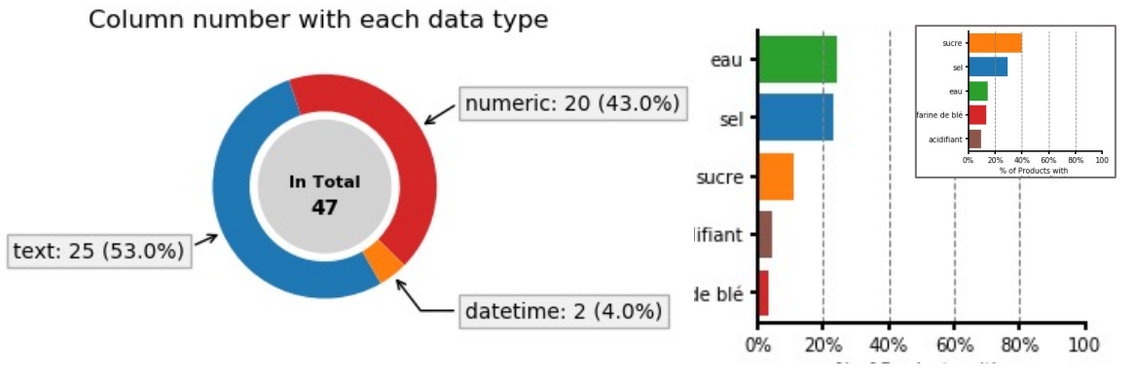
- ✓ Data type
- ✓ Value examples
- ✓ Amount of missing data
- ✓ Value ranges
- ✓ Class number

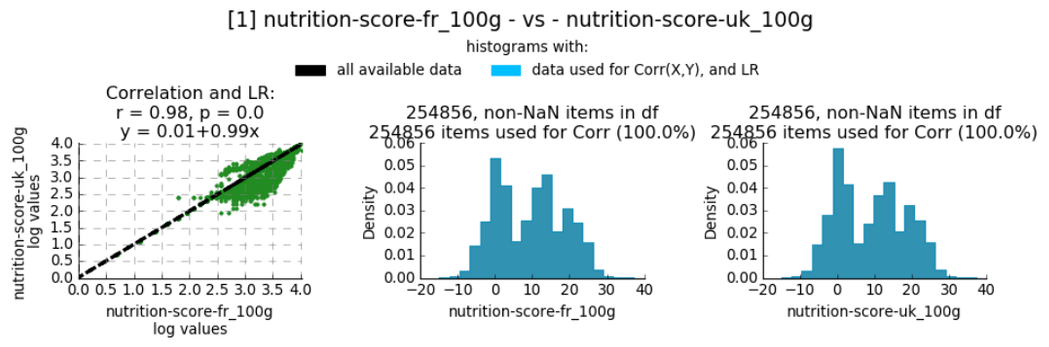


Helps finding

- ✓ Duplicates
- ✓ Correlations
- ✓ Cleaning and organizing the data
- ✓ Explore each feature







USE CASE EXAMPLE:

https://github.com/PawelRosikiewicz/HealthyFoodLabels

TM open source

Pawel Rosikiewicz

Lead Researcher on end-to-end multi-stage project

Created Entire System for production & identification of polyploid microorganisms

LABOLATORY → MOLECULAR BIOLOGY & BIOINFORMATICS → NGS DATA ANALYSIS

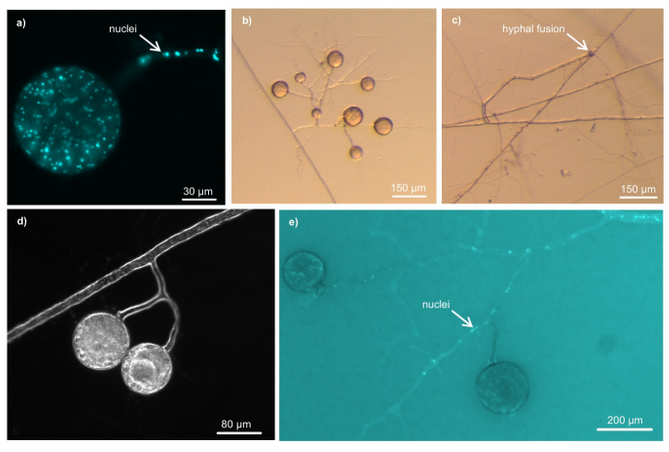
MY ROLE

- DESIGN & DEV.
- NGS SEQUENCING
- GENETICS
- MANAGED LAB TEAM
- VARIANT CALLING
- CUSTOM PIPELINES

MY PROJECT

MY MAIN EXPERIMENT

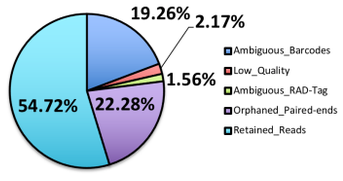
- Took 5 years
- >25.000 bio-samples
- 3TB or raw data



Sequencing

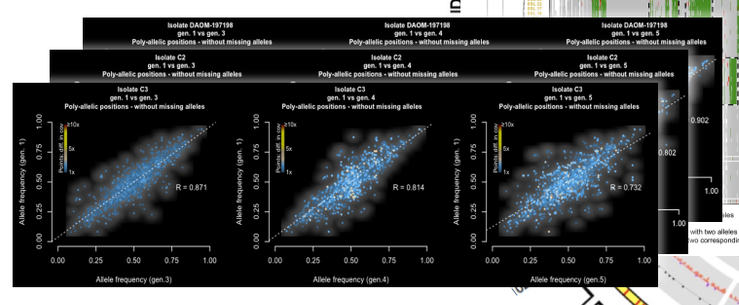
Variant Calling*

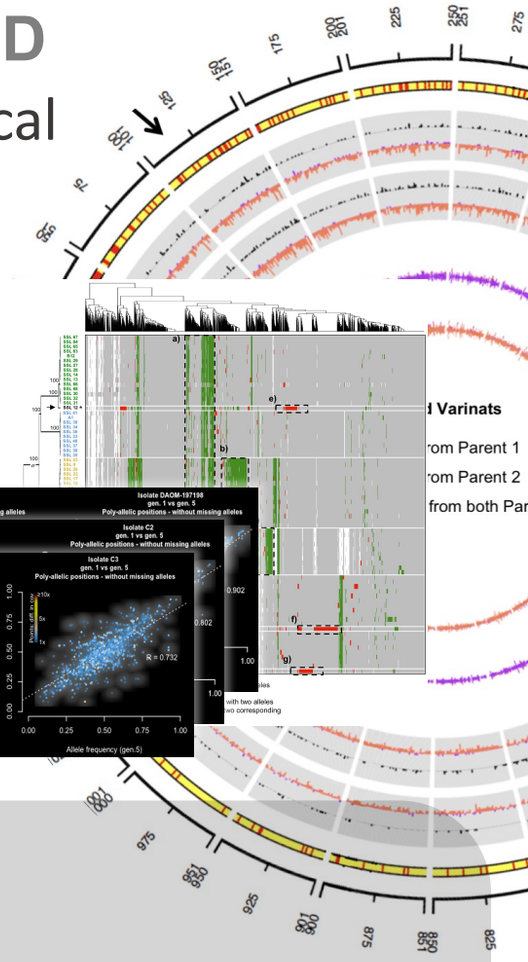
Quality Filtering



I DEVELOPED

- New Analytical Methods
- Dashboards
- Pipelines





IMAPCT

PRODUCTS GENERATED WITH MY PROJECT

IN VITRO CULTURE SYSTEM

IMPLEMENTED IN INDUSTRY Fig.1

👉 https://doi.org/10.1007/s00572-017-0763-2

NGS PROTOCOL

USED OVER 3000x

👉 https://doi.org/10.1371/journal.pone.0226497

60 STRAINS OF MICROORGANISMS

TESTED FOR USE IN AGRONOMY

👉 https://www.biorxiv.org/content/10.1101/830547v1

I used simulated data and to explain patters observed in empirical data and to build new pipelines

EXAMPLE

One of my experiments could have six possible results, and many possible subtypes

MY TASK:

- find relevant features
- create simulated data corresponding to each scenario
- Use statistical tests to find best targets
- Design new cost-efficient experiment to verify it

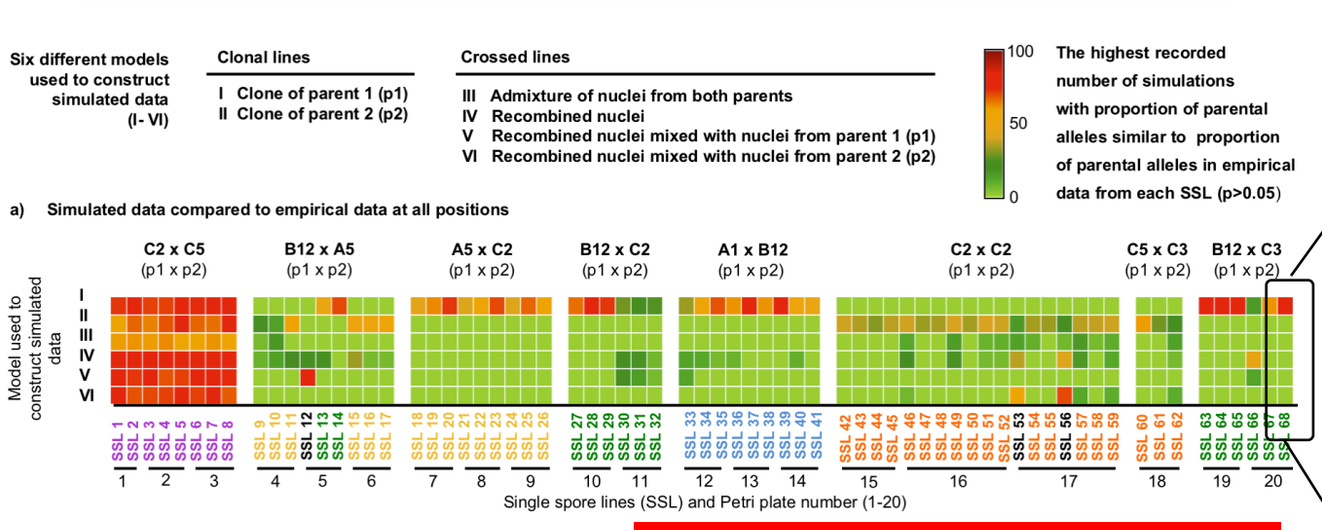
HOW?

- I used simulated data with known results to build, and test new pipelines
- I used empirical data as starting point for my simulations
- I repeated the process over 6 million times for 120.000 features, in thousands of conditions and compared them with empirical data

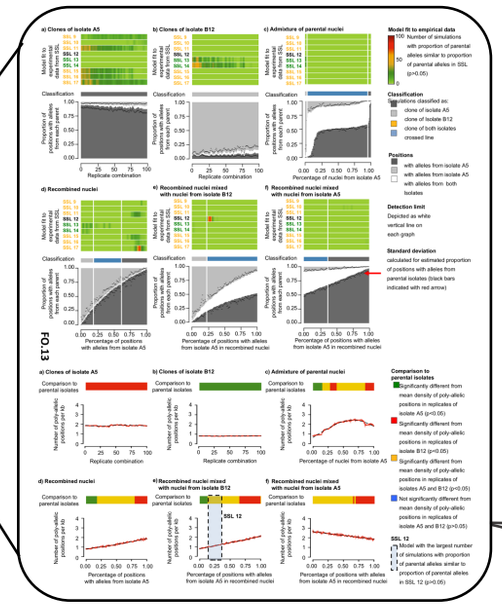
I created Dashboards to summarize the results

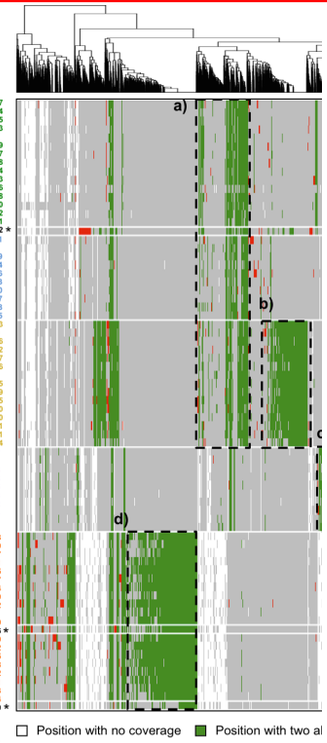
… to see more details if needed

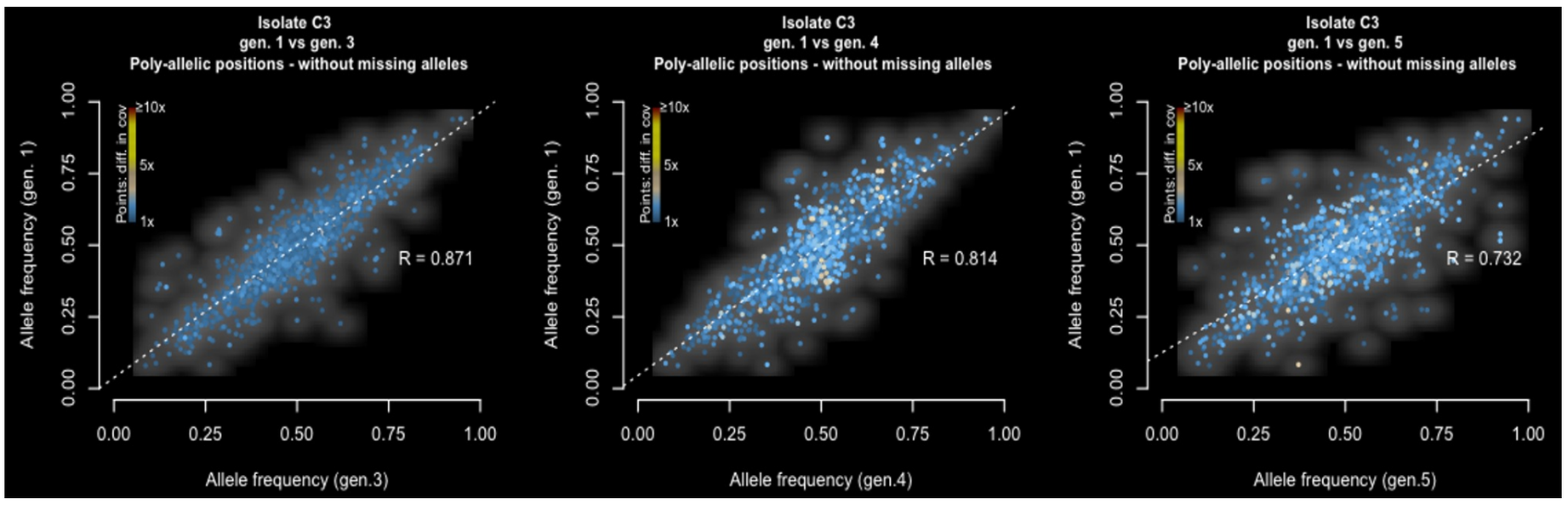
… and how samples relate to each other



+ STATS, eg acuracy, error rate, FPR etc..







or to examine each sample individually

In one year

I created a community of over 2000 AI specialists, thanks to systematic work and connection with strong brands

SwissAI

Machine Learning Meetup

Pawel Rosikiewicz

Founder and Team Leader

of SwissAI Machine Learning Meetups

one of the largest group in Europe

www.SwissAI.org